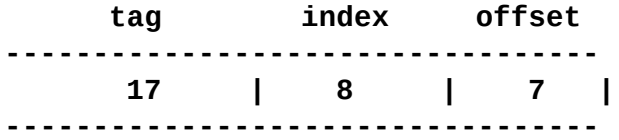
CS 3853 Midterm Exam 2 Solutions — Fall 2013

1) (4 points) $1 + .07 \times 9 = 1.63$

2) (5 points) $1 + .03 \times 17 + .28 \times .06 \times 17 = 1 + .51 + .2856 = 1.7956$

3) (5 points) Separate caches avoid a structural hazard which occurs on a load or store if there is a unified cache.

4) (4 points, 1 point each)
   a)



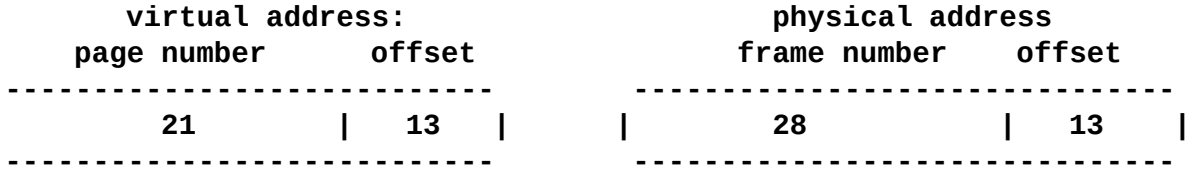
   b) as above with tag of 18 bits and index of 7 bits.
   c) as above with tag of 19 bits and index of 6 bits.
   d) as above, with no index, just a tag of 25 bits and an offset of 7 bits.

5) (2 points) 0x1b7456 = [1 1011 0111 0100 0101 0110] = [11011011101000 1010110]
   so the block address is [11011011101000] = [11 0110 1110 1000] = 0x36e8 and the offset is [1010110] = 0x56

6) (8 points, 2 points each)
   a) 0x36e8 = [11 0110 1110 1000] = [110110 11101000] so the tag is 0x36 and the set (block address) is 0xe8.
   b) The tag is [1101101] = 0x6d and the set number is [1101000] = 0x68.
   c) The tag is [11011011] = 0xdb and the set number is [101000] = 0x28.
   d) The tag is the block address: 0x36e8. There is no index, so the set number is 0.

7) (11 points)
   a) (2 points each) These describe what happens on a write hit. With write-through, any write to the cache is also done to the memory. With write-back, a write to the cache does not cause a write to memory, but the entire block is written to memory when it is replaced if it has been modified.
   b) (2 points each) These describe what happens on a write miss. With write-allocate, the cache is filled with the corresponding block, just as with a read. With no-write allocate, a write miss does not load the cache.
   c) (3 points) A dirty bit indicates that the block in the cache has been written to. It is useful when write-back is used so that a replaced block is only written back to memory if it has been modified.

8) (8 points)



   Max memory 2^{41} bytes = 2048 GB, Max process size is 2^{34} bytes = 16 GB.

9) (6 points) Access time = L1 hit time + L1 miss rate $\times$ L1 miss penalty
   L1 miss penalty = L2 access time = L2 hit time + L2 miss rate $\times$ L2 miss penalty
   L2 miss penalty = L3 access time = L3 hit time + L3 miss rate $\times$ L3 miss penalty = 17 ns + .01(50 ns) = 17.5 ns.
   L1 miss penalty = 6 ns + .07(17.5) ns = 7.225 ns.
   Access time = 1 ns + .09(7.225 ns) = 1.65025 ns.

10) (8 points, 4 points each)
    a) A page table is an array of frame numbers indexed by page numbers. A virtual address has two parts, a page number and an offset. A physical address has two parts, a frame number and and offset. The page table is used to convert the page number into a frame number and the offset is copied from the virtual address to the physical address.
    b) A TLB is a translation lookaside buffer which is a cache for the page table. It is used to speed up address translations because without it a page translation would require a memory access. An entry contains a valid bit, a page number, and a frame number.

11) (6 points, 2 points each) The offset requires 15 so a page number is 49 bits and a frame number is 28 bits. In each case a TLB entry contains a valid bit, a tag, and a frame number.
    a) The index is 5 bits, so the tag is 49-5 = 44 bits, for a total of 1(valid)+3(protection)+44+28 = 76 bits per entry or 76(32) = 2432 bits total.
    b) The index is 3 bits, so the tag is 49-3 = 46 bits, for a total of 1+3+46+28 = 78 bits per entry or 78(32) = 2496 bits total.
    c) There are no index bits so the tag is 49 bits, for a total of 1+3+49+28 = 81 bits per entry or 81(32) = 2592 bits total.

12) (13 points)

a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r
64	50	14	45	5	9	5	45	29	29	256	29	43	24	14	5	24	256

13) (10 points)
    a) direct: 42/.52 = 80.77 rounds up to 81 cycles, 2-way: 42/.59 = 71.19 rounds up to 72.
    b) direct: 1 + 1.27(.08)(81) = 9.2296, 2-way: 1 + 1.27(.07)(72) = 7.4008.
    c) direct: 9.2296(.52) = 4.79939 ns, 2-way: 7.4008(.59) = 4.36647 ns.

    The 2-way cache is better because the time per instruction is smaller.

14) (10 points)
    a) Access time is the time to access the first byte of a data access. Transfer rate is the rate at which bytes can be transferred after the first access. If each byte is accessed independently, the bandwidth is the reciprocal of the access time. For many devices (such as memory or disk), one the first byte (or word) is accessed, consecutive bytes can be accessed at a faster rate.
    b) Writing a single byte or word to disk is inefficient since an entire sector must be read or written. Write back is used so that disk accesses can be done in blocks.
    c) The L1 cache must be accessed in one cycle, and so is faster and than other on-chip caches. This makes the cache more expensive and less energy efficient per byte than the other caches. Also, if virtually indexed, physically tagged caches are used, the L1 cache size is limited by the page size.
    d) If logical addresses have many bits (such as 64) and page sizes are not overly large (such as a few KB), page numbers have a large number of bits (about 50). This would make a single-level page table very large. The added translation time is only necessary on a TLB miss, so this would be done a small fraction of the time since TLB hit ratios are very high (on the order of 99%).
    e) A cache is called virtually indexed if the cache index can be determined from the virtual address. It is called physically indexed if the physical address is needed to calculate the index. Calculating the physical address requires address translation, which at least requires access to the TLB. With virtually indexed L1 caches, the cache set can be accessed while the address is being translated, and this can reduce the cycle time, since the L1 cache should be able to be accessed in 1 cycle. For the higher level caches this is not necessary since by the time the cache is accessed, the physical address is known.